Selecting random number seeds in practice

Kleijnen, J.P.C.

Publication date:
1985

faculteit der economische wetenschappen

RESEARCH MEMORANDUM

TILBURG UNIVERSITY
DEPARTMENT OF ECONOMICS
Postbus 90153 - 5000 LE Tilburg
Netherlands

FEW
198

SELECTING RANDOM NUMBER SEEDS IN PRACTICE

Dr. Jack P.C. Kleijnen

Professor of Simulation and Information Systems
Department of Information Systems and Accountancy (ISA)
School of Business and Economics
Catholic University Tilburg (Katholieke Hogeschool Tilburg)
5000 LE Tilburg, Netherlands

september 1985

ABSTRACT

A pseudorandum number generator, like a multiplicative congruential generator, depends on its initial value or seed. The computer may select a seed using its internal clock. Alternatively the simulation analyst may use Fishman's tables with seeds spaced 100,000 apart. Problems arise if consecutive simulation runs can not be made in a single terminal session. Theoretically the concepts of sampling with and without replacement are involved. Practically the user is permitted to select the option he finds simplest. Runs for different simulated systems may use common seeds to decrease the variability; the correct analysis, however, is a controversial issue; the internal clock is an impractical source since its internal (binary) representation must be saved.

Keywords

Random number generation, statistical analysis, sampling, variance reduction.

INTRODUCTION

This note was inspired by the following practical questions. How should the seed of a pseudorandom number generator be specified, i.e., should that seed be sampled or be taken from a table with seeds "a 100,000 apart" as Fishman (1978) proposed? If the seed is to be sampled, how can this statistical concept be realized? And if Fishman's proposal is followed, then his tables should be extended to pseudorandom number generators implemented on different (micro)computers. The literature does not address the above questions explicitly, and concentrates on the design and analysis of pseudorandom generators, treating seeds as a minor detail. The present note does not pretend to present new concepts. It does try to provide practical answers; it also sketches theoretical concepts that others may wish to develop further.

THE PROBLEM

The most popular pseudorandom number generators are multiplicative congruential: in eq. (1) the symbol a denotes a (constant) multiplier, b an additive

constant, m the modulo, and these three parameters are nonnegative integers (b may be zero):

$$x_i = (a\, x_{i-1} + b) \mod m \quad (i=1,2,\ldots) \tag{1}$$

Numbers between zero and one ($0 \leq r_i < 1$) result from

$$r_i = \frac{x_i}{m}. \tag{2}$$

The literature discusses the specification of the parameters a, b and m in great detail, including statistical tests of the independence of sucessive pseudorandom numbers (ex post, empirical analysis) and (a priori, deductive) mathematical analysis (based on number theory). For a recent review see Ripley (1983); also see any handbook on simulation.

In practice the simulationist uses a computer with a given generator, i.e., the computer comes with given values for the parameters a, b and m. The user, however, does have control over the initial value or seed x_o. And he (or she) wants to make a number of "runs" with the simulation program, for example, he models a gas station as a queuing system with random arrival and service times; he wants to simulate the existing gas station during n_1 days ($n_1 > 1$); next he wants to add one more pump to the existing system and simulate its operation during n_2 days ($n_2 > 1$), and so on. How should the analyst select the seeds in his simulation experiment ($x_0^{(j)}$ in run j with j=1,...,n, and $x_0^{(k)}$ in run k with k=1,...,n_2, etc.)?

ALTERNATIVE SOURCES FOR THE SEED

The default option

The user may decide (explicitly or implicitly, i.e., knowingly or without realizing it) to let the computer decide on the seed. Usually the computer generates a seed through its internal clock, i.e., the computer looks at its "digital watch" and uses the numbers representing time to fix the seed (also see Ripley, 1984). Some (older) computers systems have a different default option, i.e., they always start with the same seed (for example,

12345678 x 2^{31}). This second variant means that all simulation experiments using the default option, are dependent. Consequently, if the experiments investigate the same system, then the conclusions are less general. Also, all experiments with the latter default option, use only part of the total pseudo-random number stream (the cycle length or period); this part may happen to have bad statistical properties. With the first option (internal clock) the bad and good parts of the cycle have the same chance, assuming the internal clock generates random seeds (a "good" random number generator also has "bad" parts, for example, hundred 6's in a row do result if a good die is thrown long enough).

External sources for seeds

The user may select a seed himself. In practice many users are lazy and pick the same simple seed (like $x_0 = 123$) in all their simulation experiments (when selecting passwords for a computer security system, many users show a similar laziness). Selecting a common seed in all simulation experiments has disadvantages, discussed above. These disadvantages can be eliminated as follows.

Fishman (1978, pp. 481-487) gives tables with seeds "spaced 100,000 apart" (say $x_0^{(1)}$, $x_0^{(2)}$,..., so that after initalizing with $x_0^{(1)}$ 100,000 calls to the generator yield $x_0^{(2)}$ and so on). If the user selects two seeds from these tables, he knows that his two simulation runs have non-overlapping streams of pseudorandom numbers. This overlap becomes of interest if several runs, each which its own seed, are considered; see the next section. First note, however, that Fishman gives these tables for only three different generators. So if the user chooses Fishman's option, he may have to construct such a table for his own specific generator (for example, he may use the generator for microcomputers proposed by Thesen, 1985).

MORE SEEDS IN ONE EXPERIMENT

The user certainly makes more than one run. The present section will show how the simulationist may select, explicitly or implicitly, the seeds of the n_1 runs with the same simulation program (in the preceding example, he simulated n_1 days of the existing gas station).

The simplest situation occurs, if the user makes these n_1 runs in a single terminal session or batch: after run 1 terminates, run 2 starts with the next pseudorandom number, for example, if run 1 stops after 500 numbers are used, then run 2 uses x_{500} as seed in eq. (1).

A different situation exists, if the user makes the n_1 runs in more than one session. There are then several sources:

(1) The internal clock (see the preceding section).

(2) Takes with seeds 100,000 apart (again see that section).

(3) A computer log, i.e., the computer saves the last pseudorandom number used in the previous session. This option has one major practical drawback, namely, internally the computer uses more bits than presented externally (on the screen or paper output). So the user cannot feed in the externally presented number as he can with Fishman's tabulated numbers. He would have to "dig" into his microcomputer or turn to the system analyst of his computer center. The differences between the options (1) and (2) are investigated in the next subsection.

Sampling with and without replacement

If the analyst uses the internal clock to specify the seed for the next run, then overlap between runs may occur. Using the tabulated seeds guarantees non-overlapping streams of pseudorandom numbers. Many authors consider this overlap as undesirable; see Fishman (1978), Mihram (1983, p. 30), Schruben and Margolin (1978, p. 507). However, theoretically sampling with replacement always means that sampling the same values is not impossible (for example, if from an urn with 1,000,000 balls a first sample of 5 balls is taken and next a second sample of 5 balls is taken, then the second sample may contain one or more individual balls of the first sample, provided the first sample was replaced). Sampling without replacement yields a certain dependence (the balls of the first sample cannot be sampled in the second sample). The difference between sampling with respectively without replacement becomes smaller, the bigger the population is. In the well-known urn example, the probability laws are known as the binomial and the hypergeometric distributions respectively;

these distributions have the same mean but the variance of the hypergeometric distribution is:

$$\text{var}\left(\sum_{i=1}^{n} x_i\right) = n\,p(1-p)\,\frac{N-n}{N-1} \qquad (3)$$

where x_i is 0 or 1 (white or red ball), n is the sample size (n=5 in the example), N is the population size (N=1,000,000) and $P(x_1=1) = p$. As N increases, eq. (3) approaches n p(1-p), the variance of the binomial distribution (sampling with replacement).

Sampling with replacement is the procedure generally advocated in statistical handbooks, since it creates independence and simplifies the statistical analysis! In the simulation literature, however, many authors advocate sampling without replacement (see the references above). In practice, the difference between the two procedures is negligible, since the pseudorandom number generator has a very long period. Therefore the user may choose the procedure he finds simplest. He probably let the computer select the seed through the internal clock, instead of using Fishman's tables or creating his own table. An exception occurs, if the user wants to apply the variance reduction techniques discussed in the next section. First, however, we add a note.

The multiplicative generator (see eqs. 1 and 2) implies that $r_i \neq r_{i'}$ (unless the parameters a, b and m are poorly chosen so that unacceptable statistical properties result) with $i \neq i'$ and $i, i' = 1, 2, \ldots, h$ where h denotes the cycle length. But this property means that sampling <u>within</u> the simulation run occurs <u>without</u> replacement. Theoretically, this property conflicts with the simulation model which specifies <u>independent</u> interarrival times. Practically speaking, the dependence created by sampling without replacement, is negligible; see the discussion of eq. (3). Moreover, if the random input variable is discrete (counterexample: exponental interarrival times) then different pseudorandom numbers may generate identical variates.

COMMON PSEUDORANDOM NUMBERS

In the gas station example, the two variants (namely the existing system and the system augmented with one more pump) may be simulated using the same pseudorandom arrival pattern of customers, i.e., using two identical seeds (for

the moment, it is assumed that arrival times are the only random component of the model). Common seeds create common pseudorandom number streams (or vector $\underset{\sim}{r}$). Common seeds tend to reduce the variance of the difference between two simulation responses: var (x-y) = var(x) + var(y) - 2 var(x,y). More generally, if many system variants are simulated, all with the same seed, then their differences tend to be estimated more accurately. Unfortunately, the statistical analysis of a simulation experiment with a common seed is controversial. Some authors analyse such an experiment using the concept of blocks; see Schruben (1979, pp. 239, 247-248) and also Anderson and Sargent (1974, p. 134), Lin and Rardin (1979, pp. 1261-1262), Schatzoff (1981, pp. 853-854). Other authors doubt the validity of the blocking model: Kleijnen (1975, p. 355), Nozari et al. (1984), Wilson (1984). And Mihram (1972, 1983) defends a different view. Recently Kleijnen (1986) proposed one more model. So common seeds create confusion, when it comes to the proper analysis.

An additional practical problem is that <u>multiple</u> seeds per run are desired. The reason is that in order to create a strong positive correlation between the responses of two or more system variants, it is advisable to use separate pseudorandom number streams per type of variable, for example, one stream for arrival times of custumers and one stream for service times of pump operators. (In simple systems a single seed suffices, for example in the gas station the pseudorandom numbers r_1, r_3, r_5,... are used for arrival times and r_2, r_4, r_6,... for service times so that no "synchronization" problem arises; see Kleijnen, 1974, p. 201.) The sources for the first K seeds (supposing K types of input variables per run; K > 1) are the same as in the situation with a single input variable (K=1), namely Fishman's tables and the internal clock. (The clock increases with one tick per pulse so that by the time the simulation program asks for the next seed, the clock has been ticking away for a "long time".) However, if the K system variants do not run in parallel, then the state of the internal clock must be saved; the externally displayed clock is inaccurate. So for all practical purposes, common seeds require externally provided seeds like Fishman's tables.

(Mihram 1983, p. 30), notes that mutiple seeds can be represented by a single seed, concatenating the K individual seeds. Antithetic pseudorandom numbers form a different variance reduction technique which, however, involves the

same issues as common random numbers do; see the references above. Common seeds are also useful in debugging, i.e., the corrected simulation program uses the same seed as the original program.

CONCLUSIONS

Statistical models are only approximations of reality. For example, the uniform distribution is used to model a "good" die and a "good" pseudorandom number generator respectively. Selecting seeds "randomly" and "100,000 apart" may be modelled as sampling with and without replacement respectively. The difference between these two models is negligible, if the population size is big, as is the case with pseudorandom numbers. Therefore practical considerations may guide the simulation analyst, for example, can the internal representation of the computer clock be saved, and do Fishman's tables apply to the generator at hand?

REFERENCES

Anderson, H.A. and R.G. Sargent (1974). Investigation into scheduling for an interactive computing system. IBM Journal of Research & Development: 125-137.

Fishman, G.S. (1978). Principles of Discrete Event Simulation. John Wiley & Sons, inc., New York.

Kleijnen, J.P.C. (1974/1975). Statistical Techniques in Simulation. Volumes I and II. Marcel Dekker, Inc., New York. (Russian translation: Publishing House "Statistics", Moscow, 1978.)

Kleijnen, J.P.C. (1986). Statistical Tools for Simulation Practioners. Marcel Dekker, Inc., New York (forthcoming).

Lin, B.W. and R.L. Rardin (1979). Controlled experimental design for statistical comparison of integer programming algorithms. Management Science, 25, no. 12: 1258-1271.

Mihram, G.A. (1972). Simulation: Statistical Foundations and Methodology. Academic Press, New York.

Mihram, G.A. (1983). Simulation methodology: statistical aspects. Proceedings 1983 Winter Simulation Conference, edited by S. Roberts, J. Banks and B. Schmeiser, published by IEEE: 27-36.

Nozari, A., S.F. Arnold and C.D. Pegden (1984). Statistical analysis under Schruben and Margolin correlation induction strategy. School of Industrial Engineering, University of Oklahoma. (Submitted for publication.)

Ripley, B.D. (1983). Computer generation of random variables - a tutorial. International Statistical Review, 51, pp. 301-319.

Schatzoff, M. (1981). Design of experiments in computer performance evaluation. IBM Journal of Research and Development, 25: 848-859.

Schruben, L.W. (1979). Designing correlation induction strategies for simulation experiments. Current Issues in Computer Simulation, edited by N.R. Adam and A. Dogramaci Academic Press, Inc., New York.

Schruben, L.W. and B.H. Margolin (1978). Pseudorandom number assignment in statistically designed simulation and distribution sampling experiments. Journal American Statistical Association. 73, no. 363: 504-525.

Thesen, A. (1985). An efficient generator of uniformly distributed random variates between zero and one. Simulation, 44, no. 1: 17-22.

Wilson, J.R. (1984). Variance reduction techniques for digital simulation. American Journal of Mathematical and Management Sciences, 4, no. 3 & 4: 277-312.

IN 1984 REEDS VERSCHENEN

138 G.J. Cuypers, J.P.C. Kleijnen en J.W.M. van Rooyen
Testing the Mean of an Asymetric Population:
Four Procedures Evaluated

139 T. Wansbeek en A. Kapteyn
Estimation in a linear model with serially correlated errors when observations are missing

140 A. Kapteyn, S. van de Geer, H. van de Stadt, T. Wansbeek
Interdependent preferences: an econometric analysis

141 W.J.H. van Groenendaal
Discrete and continuous univariate modelling

142 J.P.C. Kleijnen, P. Cremers, F. van Belle
The power of weighted and ordinary least squares with estimated unequal variances in experimental design

143 J.P.C. Kleijnen
Superefficient estimation of power functions in simulation experiments

144 P.A. Bekker, D.S.G. Pollock
Identification of linear stochastic models with covariance restrictions.

145 Max D. Merbis, Aart J. de Zeeuw
From structural form to state-space form

146 T.M. Doup and A.J.J. Talman
A new variable dimension simplicial algorithm to find equilibria on the product space of unit simplices.

147 G. van der Laan, A.J.J. Talman and L. Van der Heyden
Variable dimension algorithms for unproper labellings.

148 G.J.C.Th. van Schijndel
Dynamic firm behaviour and financial leverage clienteles

149 M. Plattel, J. Peil
The ethico-political and theoretical reconstruction of contemporary economic doctrines

150 F.J.A.M. Hoes, C.W. Vroom
Japanese Business Policy: The Cash Flow Triangle
an exercise in sociological demystification

151 T.M. Doup, G. van der Laan and A.J.J. Talman
The $(2^{n+1}-2)$-ray algorithm: a new simplicial algorithm to compute economic equilibria

IN 1984 REEDS VERSCHENEN (vervolg)

152 A.L. Hempenius, P.G.H. Mulder
Total Mortality Analysis of the Rotterdam Sample of the Kaunas-Rotterdam Intervention Study (KRIS)

153 A. Kapteyn, P. Kooreman
A disaggregated analysis of the allocation of time within the household.

154 T. Wansbeek, A. Kapteyn
Statistically and Computationally Efficient Estimation of the Gravity Model.

155 P.F.P.M. Nederstigt
Over de kosten per ziekenhuisopname en levensduurmodellen

156 B.R. Meijboom
An input-output like corporate model including multiple technologies and make-or-buy decisions

157 P. Kooreman, A. Kapteyn
Estimation of Rationed and Unrationed Household Labor Supply Functions Using Flexible Functional Forms

158 R. Heuts, J. van Lieshout
An implementation of an inventory model with stochastic lead time

159 P.A. Bekker
Comment on: Identification in the Linear Errors in Variables Model

160 P. Meys
Functies en vormen van de burgerlijke staat
Over parlementarisme, corporatisme en autoritair etatisme

161 J.P.C. Kleijnen, H.M.M.T. Denis, R.M.G. Kerckhoffs
Efficient estimation of power functions

162 H.L. Theuns
The emergence of research on third world tourism: 1945 to 1970; An introductory essay cum bibliography

163 F. Boekema, L. Verhoef
De "Grijze" sector zwart op wit
Werklozenprojecten en ondersteunende instanties in Nederland in kaart gebracht

164 G. van der Laan, A.J.J. Talman, L. Van der Heyden
Shortest paths for simplicial algorithms

165 J.H.F. Schilderinck
Interregional structure of the European Community
Part II: Interregional input-output tables of the European Community 1959, 1965, 1970 and 1975.

IN (1984) REEDS VERSCHENEN (vervolg)

166 P.J.F.G. Meulendijks
An exercise in welfare economics (I)

167 L. Elsner, M.H.C. Paardekooper
On measures of nonnormality of matrices.

IN 1985 REEDS VERSCHENEN

168 T.M. Doup, A.J.J. Talman
A continuous deformation algorithm on the product space of unit simplices

169 P.A. Bekker
A note on the identification of restricted factor loading matrices

170 J.H.M. Donders, A.M. van Nunen
Economische politiek in een twee-sectoren-model

171 L.H.M. Bosch, W.A.M. de Lange
Shift work in health care

172 B.B. van der Genugten
Asymptotic Normality of Least Squares Estimators in Autoregressive Linear Regression Models

173 R.J. de Groof
Geïsoleerde versus gecoördineerde economische politiek in een twee-regiomodel

174 G. van der Laan, A.J.J. Talman
Adjustment processes for finding economic equilibria

175 B.R. Meijboom
Horizontal mixed decomposition

176 F. van der Ploeg, A.J. de Zeeuw
Non-cooperative strategies for dynamic policy games and the problem of time inconsistency: a comment

177 B.R. Meijboom
A two-level planning procedure with respect to make-or-buy decisions, including cost allocations

178 N.J. de Beer
Voorspelprestaties van het Centraal Planbureau in de periode 1953 t/m 1980

178a N.J. de Beer
BIJLAGEN bij Voorspelprestaties van het Centraal Planbureau in de periode 1953 t/m 1980

179 R.J.M. Alessie, A. Kapteyn, W.H.J. de Freytas
De invloed van demografische factoren en inkomen op consumptieve uitgaven

180 P. Kooreman, A. Kapteyn
Estimation of a game theoretic model of household labor supply

181 A.J. de Zeeuw, A.C. Meijdam
On Expectations, Information and Dynamic Game Equilibria

182 Cristina Pennavaja
Periodization approaches of capitalist development.
A critical survey

183 J.P.C. Kleijnen, G.L.J. Kloppenburg and F.L. Meeuwsen
Testing the mean of an asymmetric population: Johnson's modified T test revisited

184 M.O. Nijkamp, A.M. van Nunen
Freia versus Vintaf, een analyse

185 A.H.M. Gerards
Homomorphisms of graphs to odd cycles

186 P. Bekker, A. Kapteyn, T. Wansbeek
Consistent sets of estimates for regressions with correlated or uncorrelated measurement errors in arbitrary subsets of all variables

187 P. Bekker, J. de Leeuw
The rank of reduced dispersion matrices

188 A.J. de Zeeuw, F. van der Ploeg
Consistency of conjectures and reactions: a critique

189 E.N. Kertzman
Belastingstructuur en privatisering

190 J.P.C. Kleijnen
Simulation with too many factors: review of random and group-screening designs

191 J.P.C. Kleijnen
A Scenario for Sequential Experimentation

192 A. Dortmans
De loonvergelijking
Afwenteling van collectieve lasten door loontrekkers?

193 R. Heuts, J. van Lieshout, K. Baken
The quality of some approximation formulas in a continuous review inventory model

194 J.P.C. Kleijnen
Analyzing simulation experiments with common random numbers

195 P.M. Kort
Optimal dynamic investment policy under financial restrictions and adjustment costs

196 A.H. van den Elzen, G. van der Laan, A.J.J. Talman
Adjustment processes for finding equilibria on the simplotope